Making Vision Transformers Truly Shift-Equivariant

Renan A. Rojas-Gomez[1] **Teck-Yian Lim**[1] **Minh N. Do**[1,2] **Raymond A. Yeh**[3]

[1]Department of Electrical and Computer Engineering, UIUC [2]VinUni-Illinois Smart Health Center, UIUC

[3]Department of Computer Science, Purdue University

CVPR JUNE 17-21, 2024 SEATTLE, WA

Motivation

Despite their notable performance on tasks like classification and segmentation, ViTs are **not** shift equivariant.

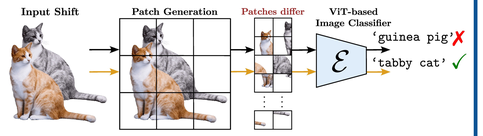

Figure 1: Effect of input shifts on ViT-based classifiers.

Goal: Design ViT modules that guarantee circular shift equivariance and enhance standard shift consistency.

- Circularly shift equivariant *by construction*.
- Generalizable across hierarchical ViT architectures.
- End-to-end trainable.

Prior Work:

- Pre-align circularly shifted images based on their polyphase representation before extracting patches.

Introduction

Modules on traditional ViTs are not shift equivariant.

- These include: patch tokenization/merging, relative position embedding and self-attention.

Result: ViT backbones break shift equivariance.

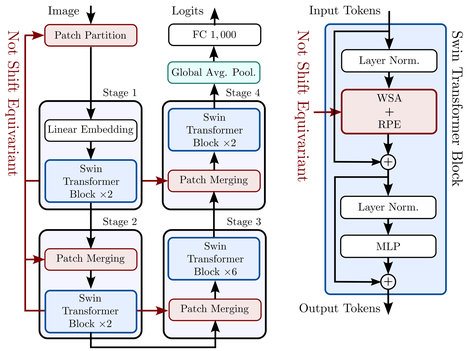

Figure 2: ViT architectures such as **Swin** are equipped with modules (highlighted in red) that break shift equivariance.

Proposed Approach

1. **Adaptive Tokenization.**

`A-token` generates patches adaptively **by maximizing a shift-invariant selection criterion** F, producing the same tokens despite circularly shifted inputs.

$$\texttt{A-token}(\boldsymbol{x}) = \boldsymbol{X}^{(m^\star)}\boldsymbol{E} \in \mathbb{R}^{\frac{N}{L}\times D},\ m^\star = \underset{0\le m\le L-1}{\arg\max}\ F(\boldsymbol{X}^{(m)}\boldsymbol{E}).$$

- $\boldsymbol{X}^{(m)} = \texttt{reshape}(\mathcal{S}_N^m\boldsymbol{x}) \in \mathbb{R}^{\frac{N}{L}\times L}$: Patch representation of the input $\boldsymbol{x}\in\mathbb{R}^N$ after circularly shifting it by m samples.
- $\boldsymbol{E}\in\mathbb{R}^{L\times D}$: Patch projection.

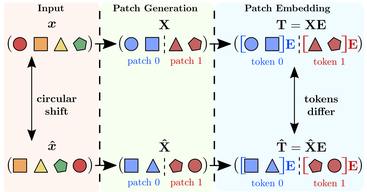

Figure 3: Original tokenization.

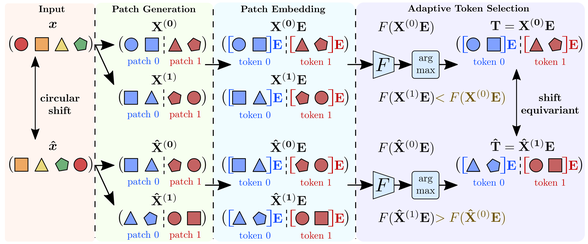

Figure 4: Proposed tokenization `A-token`.

2. **Adaptive Window-based Self-Attention.**

`A-WSA` aligns tokens via a **shift-invariant selection criterion** G to choose the window partition with the **highest energy**, producing a shift-equivariant representation.

$$\texttt{A-WSA}(\boldsymbol{T}) = \texttt{WSA}\big(\mathcal{S}_M^{m^\star}\boldsymbol{T}\big) \in \mathbb{R}^{M\times D'},\ m^\star = \underset{0\le m\le W-1}{\arg\max}\ G\big(\boldsymbol{v}_W^{(m)}\big).$$

- $\boldsymbol{v}_W^{(m)}[k] = \frac{1}{W}\sum_{l=0}^{W-1}\|(\mathcal{S}_M^m\boldsymbol{T})_{(Wk+l)\bmod M}\|_p$: Energy of the k-th window resulting from circularly shifting the input tokens $\boldsymbol{T}\in\mathbb{R}^{M\times D}$ by m indices.

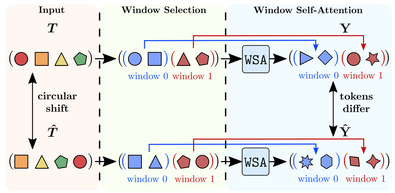

Figure 5: Original Window-based Self-Attention.

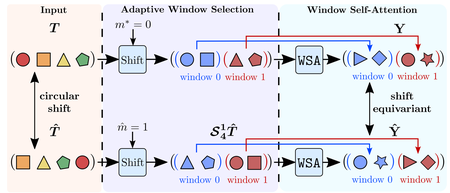

Figure 6: Proposed Window-based Self-Attention `A-WSA`.

3. **Adaptive Relative Position Embedding.**

To consider the periodicity induced by circular shifts, `A-RPE` leverages a circular relative position matrix $\boldsymbol{E}^{(\text{rel})}\in\mathbb{R}^{M\times M}$.

$$\boldsymbol{E}^{(\text{rel})}[i,j] = \boldsymbol{B}\left[(p_i^{(Q)} - p_j^{(K)}) \bmod M\right].$$

- $\boldsymbol{B}\in\mathbb{R}^{(2M-1)}$: Position embeddings.
- $p^{(Q)}, p^{(K)}$: Queries and keys indices.

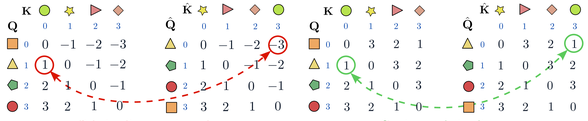

Figure 7: Original relative distance.

Figure 8: Proposed relative distance.

Results

1. **Image Classification**

Method	Circular Shift Top-1 Acc.	Circular Shift C-Cons.	Standard Shift Top-1 Acc.	Standard Shift S-Cons.
Swin-T	90.15	83.30	90.11	86.35
A-Swin-T **(Ours)**	**93.39**	**99.99**	**93.50**	**96.00**
SwinV2-T	89.08	89.16	89.08	91.68
A-SwinV2-T **(Ours)**	**91.64**	**99.99**	**91.91**	**95.81**
CvT-13	90.06	75.80	90.05	84.66
A-CvT-13 **(Ours)**	**93.87**	**100**	**93.71**	**96.47**
MViTv2-T	96.00	86.55	96.14	91.34
A-MViTv2-T **(Ours)**	**96.41**	**100**	**96.61**	**98.36**

Table 1: CIFAR-10 classification results. Top-1 accuracy and shift consistency (%) under circular and standard shifts. Bold numbers indicate improvement over the baseline architectures.

Method	Circular Shift Top-1 Acc.	Circular Shift C-Cons.	Standard Shift Top-1 Acc.	Standard Shift S-Cons.
Swin-T	78.5	86.68	81.18	92.41
A-Swin-T **(Ours)**	**79.35**	**99.98**	**81.6**	**93.24**
SwinV2-T	78.95	87.68	81.76	93.24
A-SwinV2-T **(Ours)**	**79.91**	**99.98**	**82.10**	**94.04**
CvT-13	77.01	86.87	**81.59**	92.80
A-CvT-13 **(Ours)**	**77.05**	**100**	81.48	**93.41**
MViTv2-T	77.36	90.03	82.21	93.88
A-MViTv2-T **(Ours)**	**77.46**	**100**	**82.4**	**94.08**

Table 2: ImageNet classification results. Top-1 accuracy and shift consistency (%) under circular and standard shifts. Bold numbers indicate improvement over the baseline architectures.

2. **Consistency of Tokens to Input Shifts**

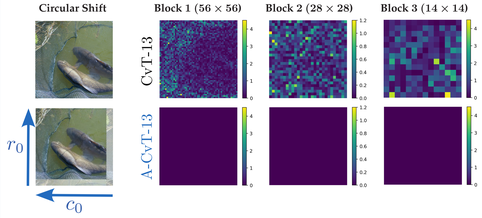

Figure 9: Consistent token representations. Small input shifts lead to large deviations (non-zero errors) in the token representations when using default ViTs (*e.g.*, CvT-13). In contrast, our proposed data-adaptive models (*e.g.*, A-CvT-13) achieve an absolute zero-error across all transformer blocks.

3. **Ablation Study**

Configuration	Top-1 Acc.	C-Cons.
A-Swin-T **(Ours)**	93.39 ± .13	**100**
No `A-token`	**93.66 ± .19**	96.29 ± .20
No `A-WSA`	93.24 ± .15	95.62 ± .54
No `A-PMerge`	91.67 ± .10	94.62 ± .11
Swin-T (Default)	90.15 ± .18	83.30 ± .61

Table 3: Ablation study. Effect of our adaptive ViT modules on classification accuracy and shift consistency (%). Configurations progressively evaluated on Swin-T under circular shifts.

4. **Semantic Segmentation**

Backbone	Circular Shift mIoU	Circular Shift mASCC	Standard Shift mIoU	Standard Shift mASSC
Swin-T	42.93	87.32	44.2	93.37
A-Swin-T **(Ours)**	**43.44**	**100**	**44.43**	**93.48**
SwinV2-T	43.86	88.16	44.26	93.23
A-SwinV2-T **(Ours)**	**44.42**	**100**	**46.11**	**93.59**

Table 4: Semantic segmentation performance. Segmentation accuracy and shift consistency (%) of our adaptive UperNet model equipped with A-SwinV2 backbones.

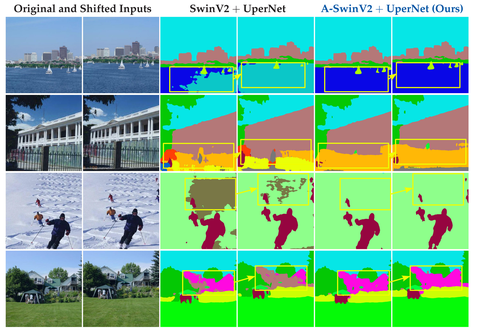

Figure 10: Semantic Segmentation under standard shifts. Segmentation results on ADE20K using our A-SwinV2 as backbone. Our adaptive model improves both segmentation accuracy and shift consistency. Examples of prediction changes due to input shifts are boxed in yellow.